Complex mutual adaptation in dyads' semantic similarity trajectories predicts conversation success

Kathryn C. O'Nell[1], Kiara L. Sanchez[1], Emily S. Finn[1]*

[1]Department of Psychological and Brain Sciences, Dartmouth College; Hanover, 03755, USA
*Corresponding author. Email: emily.s.finn@dartmouth.edu

Abstract

Conversation is a core component of human social life. Many influential theories treat conversation as a linear system, i.e., the sum of its parts. However, linear accounts of semantic entrainment lose explanatory power at their upper limits: discussing perfectly aligned opinions becomes boring. We propose an alternate account that treats conversation as a complex dynamical system. In such systems, pink noise is a hallmark of successful mutual adaptation. Here, we test whether pink noise can be found in interlocutors' co-navigated trajectories through semantic space. Not only is pink noise present, but its strength predicts positive conversation outcomes, increases with practice, and is reflected in a large language model's representation of conversation quality. These results demonstrate an important role for mutual adaptation in promoting successful conversation.

Introduction

Good conversation is like pornography – it's difficult to define, but we know it when we see it. Despite the difficulty of the task, building a mechanistic understanding of how good conversation emerges is essential to the study of human social interaction. Existing work takes several approaches: investigations based on high-level folk-psychological constructs (1-3), data-driven approaches that objectively quantify low- and mid-level features of conversation (4-7), and the application of domain-general cognitive mechanisms to the specific task of conversation (8-11). While this work has made some inroads into the thorny problem of characterizing good conversation, many of these approaches rely primarily on linear methods that assume conversation to be the sum of its parts (i.e., individual speakers and speech acts) rather than a complex, dynamical system whose behavior is irreducible to individual elements (12).

A dyadic conversation can be thought of as a system with two components: speaker A and speaker B. Like any such system, this conversation could operate in a component-dominant or interaction-dominant regime (13). In the component-dominant regime, the final state of a system can be determined simply by adding together the trajectories of the two components, with no need to consider their interactions with each other. Within a conversation, this might present itself as two individuals monologuing about their own preferred topics, with no acknowledgement of the other's presence. A less extreme, but more recognizable, example comes in the form of a conversational partner who focuses on what they intend to say next at the expense of listening and adapting. In the interaction-dominant regime, the final state of a system depends on non-linear interactions between the individual components; in other words, removing one changes the other's behavior. In natural conversation, this presents as a comfortable, mutually adaptive flow: individuals might ask follow-up questions, move between topics fluidly, or riff on each other's jokes. While this regime is often couched in the language of synchrony (e.g. "being on the same wavelength"), highly synchronous conversation does not necessarily exhibit mutual adaptation, nor does synchrony in the extreme make for good conversation – consider a scenario in which one speaker simply parrots exactly what the other says. This conversation would be both perfectly synchronous and extraordinarily boring.

An important hallmark of interaction-dominant systems is that they exhibit pink noise: a signal where power is inversely proportional to frequency on a log-log scale, such that lower frequencies show higher power (14, 15). These signals sit halfway between purely stochastic white noise, with equal power in all frequencies, and Brownian drift, where lower frequencies dominate. While the origins of this phenomenon are not fully understood, one hypothesis is that pink noise is produced by self-organizing dynamical systems that persist at the edge of chaos – balancing stability with the flexibility to evolve to vastly different states with minimal perturbations (16). Pink noise appears in a wide range of healthy coordination behaviors: postural sway (17), heart rate variability (18), and EEG oscillations (19). It also appears specifically in adaptive coordination during interpersonal interactions: dyads with loosely-coupled body movements perform better than their uncoupled and tightly-coupled counterparts at joint problem-solving tasks (20), and for children on the autism spectrum, those with better social function have stronger pink noise signals in their eye contact onset and offset patterns during conversation (21). However, the bulk of the work connecting pink noise with interpersonal coordination has focused on paralinguistic signals (e.g., body movements, eye

contact), and not words themselves; a focus on semantics, rather than syntactic or lexical features (22-27), is even more rare (28, 29). But if we wish to understand conversation, not just interaction in general, semantics are the primary channel of interest. When asked to recount a conversation, few people would think to mention their partner's postural sway or prosodic cues. Instead, we share the topics that were discussed: what the conversation was about.

Here, we demonstrate that interlocutors' interaction-dominant negotiations in semantic space, as operationalized by pink noise, predict positive conversational outcomes like enjoyment and connection. We find evidence for this in two independent dyadic conversation datasets, in addition to naturally occurring sources of data like two-host podcasts and speed-dating encounters. We also demonstrate that large language models asked to produce enjoyable conversations produce scripts with stronger pink noise signals, suggesting that this relationship between mutual adaptation and enjoyment is encoded in the large quantity of human-generated text used to train large language models. Our results confirm theoretical predictions about the role of semantic mutual adaptation in conversation, and position complexity-based approaches as a viable framework for mechanistically understanding conversational success.

Materials and Methods

Generating a conversation's semantic similarity trajectory

All conversations analyzed in this paper first took place in the spoken modality and were then transcribed into text by various methods. For each conversation transcript the text was split into turns, defined as any words contained in one speaker's uninterrupted flow of speech (Figure 1A). Each turn was then passed through MPNet, a transformer model that analyzes how words relate to and influence each other within a sentence to capture its overall meaning (39). MPNet converts each turn into a numerical representation: a vector of 786 numbers that defines that turn's location within a high-dimensional semantic space where similar meanings cluster together (Figure 1B, top). The cosine distance was calculated then between each turn's embedding and that of the prior turn. For a conversation of length n, this produces a length n-1 signal representing interlocutors' movements towards and away from each other in semantic space (Figure 1B, bottom).

Identifying pink noise within semantic synchrony trajectories

We applied Detrended Fluctuation Analysis (DFA) to each conversation's semantic similarity trajectory to determine the signal's noise scaling coefficient (30, 31). Briefly, DFA analyzes how fluctuations in a signal scale across different timescales. It works by (1) creating a cumulative sum of the time series, (2) dividing this into segments of length n, (3) detrending each segment by subtracting its linear fit, (4) calculating the root-mean-square fluctuation around zero for each segment length n, and (5) examining how these fluctuations scale with different values of n. The scaling relationship reveals whether the signal is dominated by short-range randomness or exhibits long-range temporal correlations. Scaling coefficients near 0.5 indicate white noise, coefficients near 1 represent pink noise, and coefficients near 1.5 represent red noise, or Brownian drift.

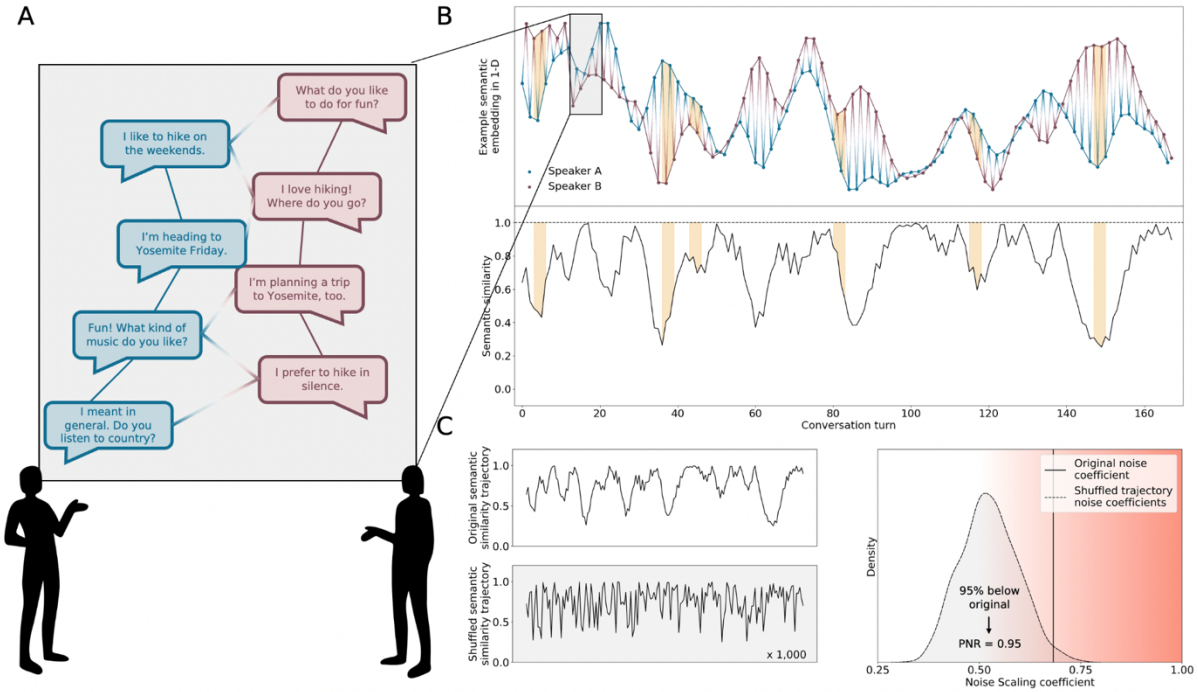

Fig. 1. Methods overview. (A) An artificial conversation snippet. (B) A depiction of how a semantic similarity trajectory is generated for a single conversation transcript. Top: each point represents the one-dimensional projection of a 768-dimensional semantic embedding of a conversational turn: see the expanded box for an example. The red and blue lines represent each of the two speakers' separate trajectories through semantic space, and the gradient lines connecting each adjacent turn represent the (cosine) distance between these turns' embeddings. Bottom: a depiction of the whole conversation's semantic similarity trajectory. Note that the yellow bars connecting points on this trajectory to the maximum value of 1 are the same length as the yellow bars in the top of this figure, which represent semantic distance between embeddings of adjacent turns. (C) A depiction of how a semantic similarity trajectory's pink noise robustness (PNR) is calculated. Left top: the example semantic similarity trajectory from B, which has a noise scaling coefficient of 0.71. Left bottom: an example similarity trajectory generated by shuffling the signal in the panel above. A null distribution of noise scaling coefficients is built for each conversation by repeating this shuffling process 1,000 times and calculating a noise scaling coefficient for each of these shuffled trajectories. Right: a depiction of a conversation's original noise scaling coefficient as compared to its null distribution. A conversation's pink noise robustness (PNR) is calculated as 1 - the proportion of shuffled trajectory noise coefficients that are pinker than the original.

To control for systematic biases in scaling coefficients introduced by conversation length, we generated a conversation-specific null distribution of noise scaling coefficients by scrambling each conversation's semantic similarity trajectory 1,000 times. We then calculated the percentage of scrambled-trajectory scaling coefficients that were below the original-order conversation scaling coefficient, a measure which we call pink noise robustness (PNR; Figure 1C).

CANDOR corpus

The CANDOR corpus is a large, open dataset consisting of 1,656 video-chat conversations in English between participants recruited through Prolific (5). Data collection and experimental procedures were approved by Ethical & Independent Review Services, protocol #19160-0, and all participants provided informed consent both before and after taking part in the study. Participants were instructed to talk as if they had just met at a social event, for at least 25 minutes (mean length: 31 minutes; SD: 7.96 minutes). We used the Cliffhanger-generated transcripts provided with the corpus to calculate pink noise robustness (PNR) scores for each conversation.

Before and after each conversation, participants took extensive surveys about their personalities and experiences. We separated the 205 survey questions that resulted in numerical responses into 6 categories: those relating to (1) conversation enjoyment, (2) sense of ongoing connection, (3) engagement with and memory for the conversation, (4) demographics and low-level conversation statistics (like the participant's perception of the average turn length and which partner smiled more), (5) how each participant rated their partner on various trait batteries, and (6) how each participant scored on those same trait scales themselves.

We ran two sets of analyses to identify relationships between a conversation's PNR and the interlocutors' survey results. First, we correlated a conversation's PNR with each of the 205 numerical survey values (Pearson correlation for continuous variables, Spearman for Likert-scaled variables) which were averaged between the two speakers for a given conversation, meaning that each conversation contributed one datapoint to this analysis. We corrected for multiple comparisons in two ways: first with a conservative Bonferroni correction, then with the more lenient Benjamini/Hochberg False Discovery Rate (FDR). After finding the sets of variables that survived the two different multiple hypothesis correction methods, we tested whether enjoyment- or connection-related variables were overrepresented in these sets using Fisher's exact test. Additionally, we took the first principal component of the survey responses within each of our 6 question categories (replacing any missing values with the median value of the variable) and correlated them with PNR.

We also performed two control analyses to confirm that conversation success was best predicted by mutual adaptation between interlocutors and not by features of the individuals themselves. First, we modeled the first principal components of individuals' survey answers in the enjoyment and connection categories as a linear function of the conversation-level PNR, the PNR of a half-length conversation made of just their turns, and the PNR of a half-length conversation made of just their partner's turns, with no random effects. Then, we isolated the 94 participants who took part in five or more conversations, who participated in 1,041 conversations total, and tested whether PNR scores were more similar within-individual than across-individual by comparing the pairwise distances in PNRs to pairwise difference in identity using a Mantel test.
We performed an additional control analysis to confirm that PNR could explain variance in conversation outcomes beyond that attributable to simpler features of the semantic similarity trajectory. For the enjoyment and connection categories, we modeled the first principal component of conversation-level outcome variables as a linear function of PNR and the mean, slope, and variance of a given conversation's semantic similarity trajectory. The input variables were all z-scored using the built-in R *scale* function to generate interpretable effect sizes.

<u>Interracial Friends corpus</u>

The Interracial Friends corpus consists of 97 video-chat conversations between Black and non-black friends recruited on college campuses who spoke for around 20 minutes (mean length: 18.0 min, SD: 9.3 min) about a formative experience that the Black friend shared at the beginning of the conversation. Data collection and experimental procedures were approved by Stanford IRB protocol # 41090, and all participants provided informed consent before taking part in the study. All conversations were transcribed by professionals hired through Scribie. These participants also answered a battery of pre- and post-conversation surveys from which we identified three sets of questions that captured conversational enjoyment, sense of connection, and individual participant traits. We correlated the first principal component of each of these categories with PNR across conversations.

Speed dating corpora

The speed dating corpora consist of episodes from two speed-dating podcasts: Blink Date and Kings and Kweens.

After excluding conversations with fewer than 20 turns, the Blink Date corpus consists of 31 10-minute-long audio-only phone conversations between two strangers. All but one of these conversations was transcribed by the podcast hosts, and the final transcript was produced using WhisperX and diarized with NeMo (40). If at the end of the conversation both parties said "yes" or "maybe" to a second date, it was considered a match. We tested whether conversations that resulted in matches had higher average PNR values than those that did not result in matches with a one-tailed t-test.

The Kings and Kweens corpus consists of 52 episodes of a speed dating podcast in which one person (the "monarch") goes on three, or occasionally four, consecutive in-person 10-minute speed dates in front of a live audience at a bar, then decides whom they would like to ask on a second date. (Note: episodes resulting in zero or more than one second date requests were excluded from this corpus.) All conversations were transcribed with WhisperX and diarized with NeMo (40) To analyze the relationship between PNR and interest in a second date, we assigned each monarch-date pair one of 3 ordinal scores: highest PNR in the episode, lowest PNR in the episode, and middle PNR. For episodes with four dates, both the second and third highest PNR conversations were given the middle PNR label. We then modeled the outcome of each date as a function of their PNR order in a logistic GLM with no random effects.

Longitudinal conversation corpus

The longitudinal conversation corpus consists of repeated conversations between a set of 25 dyads collected from podcasts. To be included in this corpus, a podcast had to meet a predetermined set of criteria regarding access, length, structure, quality, amateurism, and familiarity. Specifically, each podcast had to be accessible via a public RSS feed, consist of at least 50 episodes, have an average episode length of over 30 minutes consist primarily of free-flowing conversations between two hosts (occasional guest episodes were acceptable, but formats in which hosts read text from social media posts or separated their conversations into strict segments were not), have no or minimal interruptions from advertisements, have positive

reviews when available, be recorded by amateurs (which we defined as anyone who did not primarily make their living from podcasting, hosting radio, or acting at the time of starting their podcast) who do not otherwise speak to each other every day (e.g., married couples). After finding 25 podcasts that fit these criteria, we removed guest episodes, live episodes, and sponsored episodes (e.g., some podcasts allow paid subscribers to request specific topics). This left 1,032 usable conversations which were then transcribed with WhisperX and diarized with NeMo (40). Finally, we modeled conversation PNR as a linear function of episode number, with a random effect of podcast.

Large language model generated conversations

We asked a large language model, Claude 3.7 Sonnet (33), to write 50 high- and low-enjoyment dialogues between strangers who have just met at a coffee shop, each with 100 turns. These conversations were generated using the following prompt:

> <system>Hi Claude! I'm running an experiment where I want to test the intuitions that you've developed by reading tons of human-generated text about what makes a conversation between two people feel like they're building a connection and enjoying each other's company — essentially what makes a conversation good. However, I don't want you to explicitly tell me about your intuitions, I want you to write me some example good and bad conversations! Please write conversations that sound as natural and realistic as possible within the provided constraints. As a note, I've noticed that you have a tendency to mirror my writing style — please don't do that here! Again, the purpose is to test the intuitions and representations that you've developed by reading lots of other people's text. </system>
> <instructions>
> <item> I'm going to ask you to write 100 conversation scripts, 50 good conversations and 50 bad conversations.</item>
> <item> Each conversation will take place between two strangers who have just met at a coffee shop.</item>
> <item> These conversations should all be distinct from each other: do not repeat phrasing or structure, and the topics should be diverse. <thinking> Take a minute to think about 50 uniquely good conversations between strangers, and 50 uniquely bad conversations between strangers. Be creative!</thinking></item>
> <item> Each conversation will consist of 100 turns, consisting of 50 back-and-forths labeled A1, B1, A2, B2 ... etc until A50 and B50. </item>
> <item> Do not add any additional speakers: these conversations are strictly between two people. </item>
> <item> Do not use action tags describing what each person is doing. </item>
> <item> Make sure that each turn contains at least one word of reply - no silences. </item>
> <item> Save each of these conversation scripts as its own markdown artifact with a title that indicates condition (good or bad). When creating the markdown filenames, please add "no-specified-length" somewhere in the title.
> </instructions>
> <thinking> Please think about these and ask clarifying questions about anything you need! </thinking>

To confirm that third-party human readers agreed with Claude's assessment of conversations' enjoyability, we asked at least 5 online participants recruited with Prolific to read each conversation and rate how enjoyable the dialogue was on a scale of 1 to 5. Data collection and experimental procedures were approved by the Committee for the Protection of Human Subjects at Dartmouth College under Protocol #00032009, and all participants provided informed consent before taking part in the study. After confirming that human raters agreed with Claude's assessment, we compared the PNRs of the high- and low-enjoyment conversations with an independent-samples t-test.

Results

Pink noise emerges in mutual semantic adaptation during conversation

Our first goal was a proof of concept: to demonstrate that pink noise, as an index of the interaction-dominance of interlocutors' movement towards and away from each other in semantic space, is present in conversation. For each conversation in the CANDOR corpus (5), an open dataset consisting of 1,656 open-ended conversations between strangers, the transcript was split into turns, embedded in high-dimensional semantic space, and the distance between adjacent turns was calculated to produce a semantic similarity trajectory for the conversation (Fig. 1; see Methods for more details). For each semantic similarity trajectory, detrended fluctuation analysis (30, 31) was used to determine a noise scaling coefficient for the conversation. A coefficient of 0.5 represents white noise, while a coefficient of 1.0 represents pink noise. Within the CANDOR corpus, 94% of conversations had noise scaling coefficients greater than 0.5 (mean=0.62, SD=0.08), suggesting that these conversations exhibit some degree of interaction-dominant dynamics in their semantic similarities trajectories.

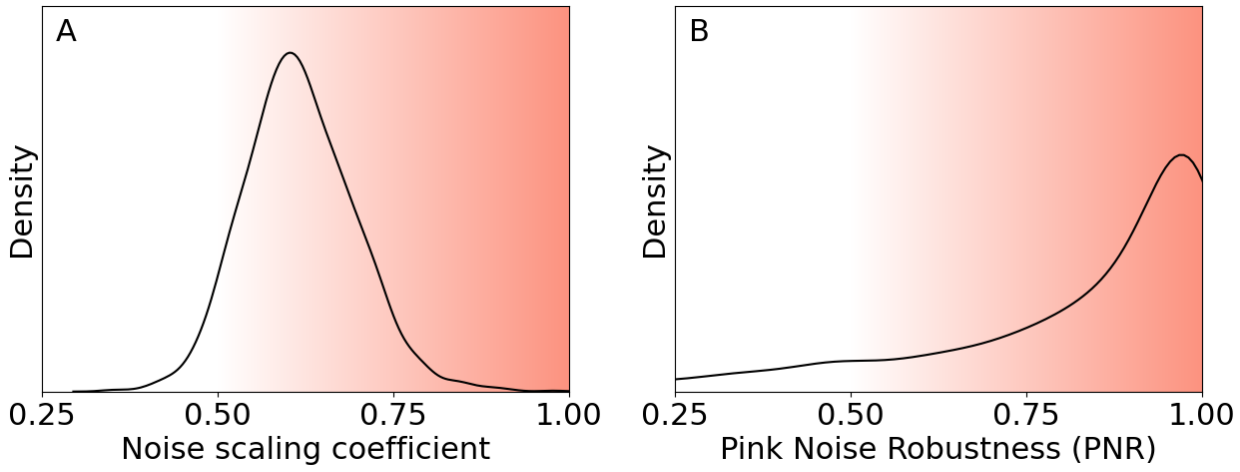

Fig 2. Distribution of noise scaling coefficients and pink noise robustness scores for the CANDOR corpus. (A) For semantic similarity trajectories of conversations in the CANDOR corpus (N=1,656), noise scaling coefficients fell between 0.35 and 0.98, with a mean of 0.62 (SD=0.08), indicating the presence of pink noise. (The scaling coefficient for pure while noise is 0.5, while the scaling coefficient for pure pink noise is 1.0). (B) For semantic similarity trajectories in the CANDOR corpus, the mean PNR score was 0.82 (SD=0.22), indicating that t the pink noise present in these conversations generally exceeds that predicted by conversation-level null distributions.

While signal noise color is a size-invariant measure, two complications can arise when working with short (length ~300 or shorter) signals. First, fewer opportunities to observe power at low frequency bands can result in systematically lower (whiter) noise scaling coefficients. On the other hand, short stochastic signals produce a broader distribution of scaling coefficients than their longer counterparts, raising the likelihood of a spuriously high (pinker) scaling coefficient. To account for these two sources of bias, we transformed the raw noise scaling coefficients into pink noise robustness (PNR) scores by comparing them to a conversation-specific null generated by calculating the scaling coefficients for 1,000 shuffled semantic similarity trajectories (see Methods for more detail). A PNR score of 0.9, for example, indicates that the observed conversation demonstrates pinker noise in its semantic similarity trajectory than 90% of shuffled versions of the same trajectory. The CANDOR conversations had a mean PNR score of 0.82 (SD=0.22), confirming that observed noise scaling coefficients above 0.5 were not merely artifacts of signal length. While the conversations demonstrate evidence of interaction-dominance en masse, there is still considerable variability in PNR scores across conversations. This variation is useful, however, because it allows us to test whether the strength of the pink noise signal in a conversation predicts positive conversational outcomes.

Pink noise predicts enjoyment and sense of connection during conversation

Having established that movement through semantic space in CANDOR conversations demonstrated varying degrees of pink noise, we then sought to test whether the strength of these pink noise signals was associated with positive conversation outcomes. In the CANDOR dataset, participants in each conversation answered extensive pre- and post- conversation surveys. We sorted the 205 variables that elicited numerical responses into six categories: those related to enjoyment, connection, engagement and memory, demographic information, one's own personality traits, and one's perception of their partner's personality traits.

Using these survey results, we took two approaches to relating pink noise to conversation success. First, in a mass univariate approach, we correlated PNR scores with each of these 205 variables across conversations. After Bonferroni multiple hypothesis correction, seven outcome variables were significantly positively correlated with PNR: how enjoyable participants found the conversation, the degree to which an interlocutor believed their partner found them friendly, how much a speaker disclosed, how much a participant believed their partner disclosed, a participant's consistent desire to remain in the conversation rather than ending early, how much longer a participant believed their partner would have liked to keep talking, and how long the conversation lasted beyond the required 25 minutes. With a more lenient Benjamini/Hochberg false-discovery rate multiple hypothesis correction, the number of outcome variables significantly correlated with PNR increased to 46, 19 of which were enjoyment-related, and 9 of which were connection-related (Fig. 3A). Both enjoyment- and connection-related variables were vastly overrepresented in the sets that survived multiple hypothesis correction (Bonferroni version: $p<0.001$; FDR version: $p<0.001$).

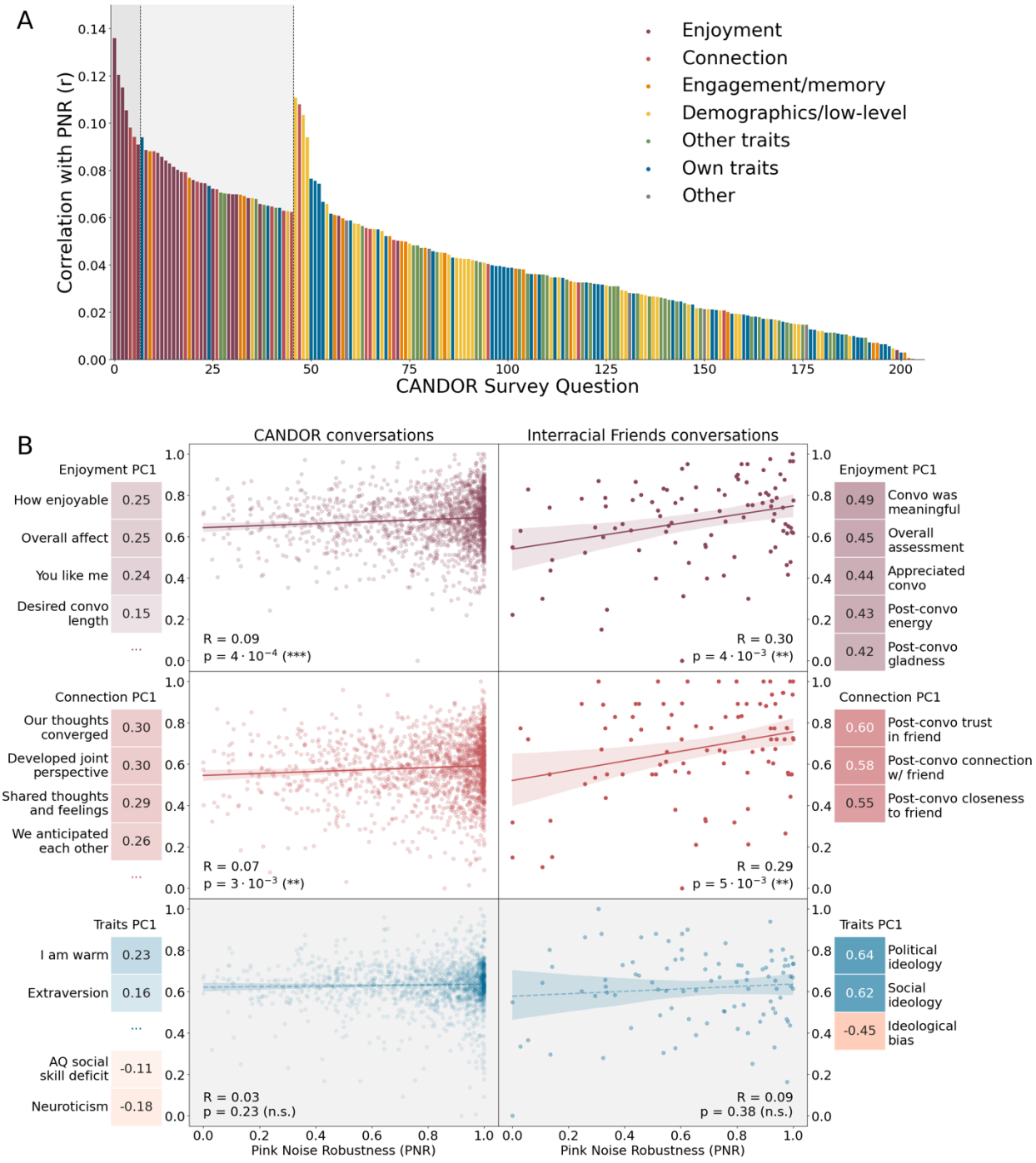

Fig 3. Pink noise robustness predicts positive conversation outcomes in two corpora. (A) Results from a mass-univariate analysis correlating pink noise robustness scores with each of the 205 numerical variables in the CANDOR survey across the approximately 1,656 conversations in the dataset. Variables are sorted along the x-axis by statistical significance (p-value), from more significant (left) to less significant (right) and colored by question category. After Bonferonni correction for multiple comparisons, seven variables were significantly correlated with PNR (dark grey background), and after Benjamini/Hochberg false-discovery rate correction, 49 variables were significantly correlated with PNR (light grey background). Enjoyment- and connection-related variables were vastly

over-represented in these sets of variables (Bonferroni version: $p<0.001$; FDR version: $p<0.001$). Note that some of the survey variables that did not survive multiple hypothesis correction have higher correlation coefficients than those with significant correlations; this is because different questions in the CANDOR survey had different response rates, leading to different degrees of freedom across the correlations. (B) PNR is correlated with the first principal component of enjoyment (top) and connection (middle) related variables, but not individual traits (bottom) in both CANDOR (left) and Interracial Friends (right) conversations. While each category (enjoyment, connection, and individual traits) is constructed from different sets of questions in the CANDOR and Interracial Friends corpora, they capture comparable conceptual dimensions. Significance is indicated as following: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Given that many survey variables were highly collinear with one another, especially within categories, we additionally tested whether conversation PNR was correlated with a summary measure of each category (i.e., score on the first principal component from a principal component analysis performed on all variables from that category). First principal component scores for enjoyment and connection were significantly correlated with PNR after multiple hypothesis correction ($R=0.09$ and 0.07; $p<0.01$ and 0.01; Fig. 3B, left), while this was not true for the rest of the categories (engagement/memory: $R=0.05$, $p=0.05$; demographics: $R=0.05$, $p=0.04$; own traits: $R=0.03$, $p=0.23$; partner's traits: $R=0.04$, $p=0.11$).

This pattern of results suggests that a stronger presence of mutual adaptation in a conversation's semantic similarity trajectory, as indexed by a higher dyadic PNR score, predicts positive outcomes like enjoyment and connection. However, this does not necessarily prove that interaction-dependent dynamics explain variance in positive outcomes beyond that explainable by a component-dominant account. If positive conversation outcomes are due primarily to cooperative mutual adaptation between interlocutors (the interaction), we should not be able to predict these same outcomes based on individual speakers' decontextualized behavior (the components). In a control analysis, we separated conversational turns into those from each speaker, then calculated "individual PNR" scores for each individual's set of turns. In a multiple regression model predicting conversation enjoyment based on dyadic PNR, a participant's individual PNR, and their partner's individual PNR, all three were significant predictors, with the largest effect coming from dyadic PNR (dyadic PNR: coefficient=0.96, $p<0.001$; participant PNR: coefficient=0.44, $p=0.03$; partner's PNR: coefficient=0.78, $p<0.001$). In an identical model predicting sense of connection, only dyadic PNR was significantly predictive (dyadic PNR: coefficient=0.66, $p<0.01$; participant PNR: coefficient=0.15, $p=0.35$; partner's PNR: coefficient=0.27, $p=0.10$). Note that this analysis was particularly conservative because an individual's semantic trajectory cannot be fully disentangled from the mutually adaptive context in which it was generated. Nevertheless, dyadic PNR provided significant additional predictive power beyond individual PNRs for both enjoyment and connection. As an additional confirmation that the predictive power of PNR was not reducible to individual behavior, we tested whether PNR scores for interlocutors who participated in many conversations were more similar within-individual than across-individual, which they were not ($R=0.001$, $p=0.12$). This suggests that, while individual interlocutors may have distinct styles of conversation, one participant's behavior alone is not necessarily enough to drive consistent mutual adaptation across conversations.

A similarly parsimonious account of conversational success comes from the application of linear methods, rather than a complexity-based approach, to analyzing a conversation's semantic similarity trajectory. While interaction-centric approaches like our pink noise analysis capture

temporal patterns in variability, linear approaches are limited to either static summary statistics that miss variation, or variance measures that obscure temporal trends. So in a second set of control analyses, we tested whether outcomes could be predicted from the average semantic similarity between pairs of adjacent turns across the whole conversation (a static measure), the variance in similarity across pairs of turns (a close inverse analogue of time-locked covariation in semantic space: a value of 0 would indicate that speakers kept a constant, rigid distance in semantic space between them), and the slope of the similarity trajectory over time (simple synchronization, without a role for adaptive desynchronization). In a multiple regression model predicting conversation enjoyment based on these three variables and PNR, all but average similarity were significant predictors (PNR: coefficient=0.25, $p<0.01$; average similarity: coefficient=-0.13, $p=0.30$; similarity variance: coefficient=-0.51, $p<0.001$, similarity slope: coefficient=0.22, $p=0.01$). In an identical model predicting sense of connection, only PNR and similarity variance were significant predictors (PNR: coefficient=0.16, $p=0.01$; average similarity: coefficient=-0.88, $p=0.37$; similarity variance: coefficient=-0.43, $p<0.001$, similarity slope: coefficient=0.12, $p=0.07$). These results suggest that complex mutual adaptation in semantic space provides insights into positive conversational outcomes that go beyond those provided by linear approaches.

Replication: pink noise predicts enjoyment and sense of connection in conversations about difficult topics between friends, not just small talk between strangers

While the CANDOR corpus is large, it consists only of a specific type of conversation: casual small talk between strangers. To test whether pink noise robustly predicts positive outcomes across different conversational settings, we conducted a conceptual replication in a separate conversation dataset with different properties: the Interracial Friends corpus. Instead of small talk, interlocutors share formative experiences; instead of strangers, these conversations occur between friends. As in the CANDOR corpus, participants answered post-conversation survey questions asking about their experience of the conversation (i.e., enjoyment, sense of connection), as well as more general questions about demographics, traits, and predictions about their future relationships. In the Interracial Friends conversations, PNR still significantly predicts both enjoyment ($R=0.30$, $p<0.01$) and connection ($R=0.29$, $p<0.01$), but is not correlated with a speaker's individual traits ($R=0.09$, $p=0.38$; Fig. 3B, right). This result suggests that mutual adaptation in semantic space is a useful framework for understanding positive conversation outcomes across different contexts and types of conversations.

Pink noise predicts real-world behavioral outcomes during speed dates

While surveys are an undeniably useful tool for assaying conversationalists' experiences, they are also subject to individual differences in scale use, motivational factors, and participants' perceptions of researchers' goals. To mitigate these effects, we sought to test whether PNR could predict positive conversational outcomes outside of a lab setting. Unlike many real-world conversations, which have only indirect or unmeasurable outcomes, first dates result in a binary signal of success: whether the pair pursues a second date. We used conversations from two speed-dating podcasts to test whether PNR could predict which couples would choose to see each other again.

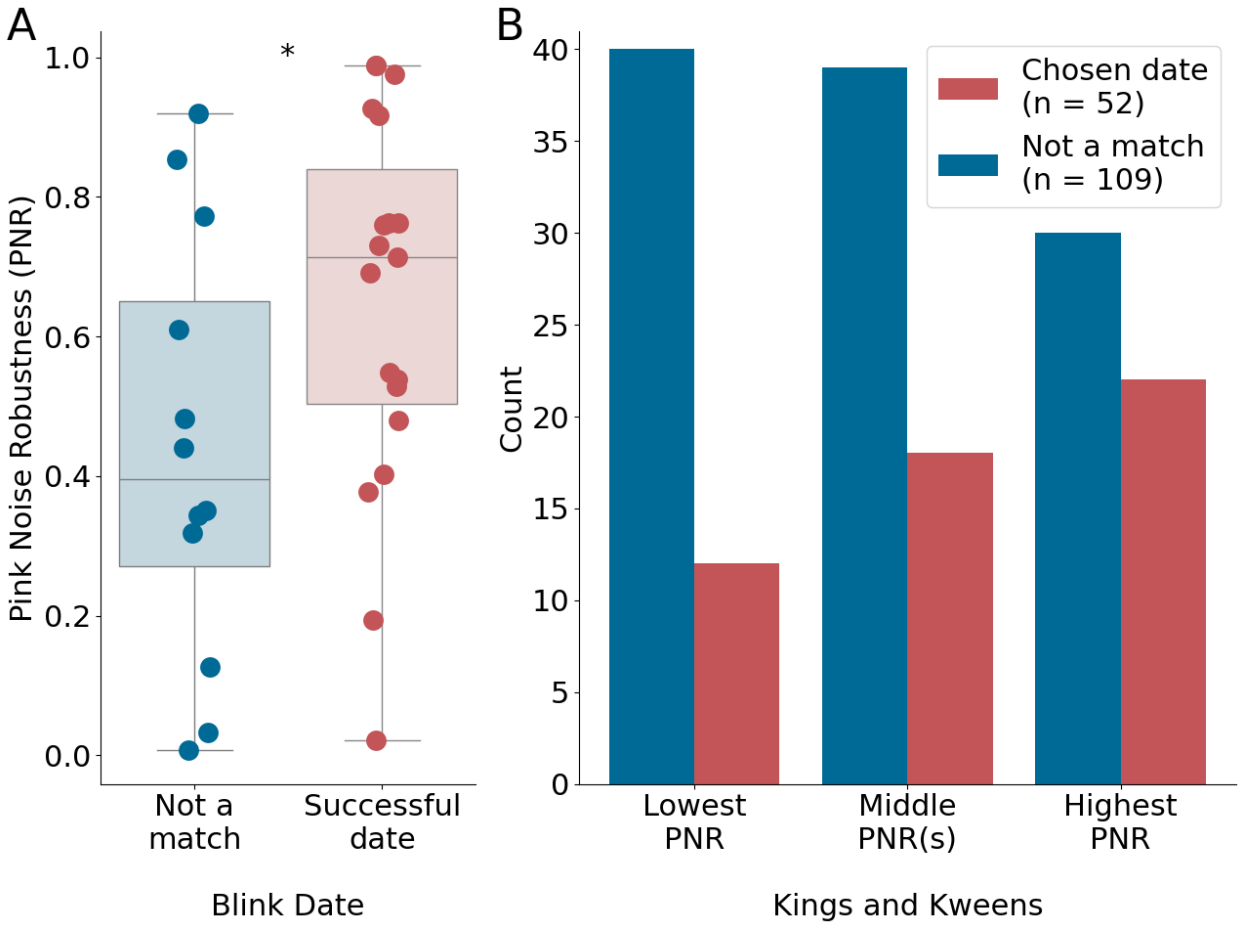

Fig. 4. Relationship between pink noise robustness and partner choice in real-world speed dates. (A) The distribution of PNRs for successful and unsuccessful first dates from the Blink Date corpus. Successful dates exhibit stronger pink noise signals than unsuccessful ones. (B) Counts of how many (un)successful dates that exhibited the lowest, middle, or highest PNR within an episode of Kings and Kweens. The likelihood of a conversation resulting in a second date increases as the PNR rank increases. Significance is indicated as following: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In the Blink Date corpus, pairs engaged in 10-minute phone conversations prompted by a light-hearted question (e.g. "What is your most controversial food opinion?"). After the date, both participants privately indicated whether they would like to go on a second date with their partner and were considered a match if both indicated "yes" or "maybe." In this dataset, successful dates had significantly higher PNRs than those that did not result in matches (t-stat=2.0, p=0.03; Fig. 4A).

Dates in the Kings and Kweens corpus were organized differently – in this podcast, a single "monarch" engages in three (occasionally four) 10-minute dates in a row in front of a live audience. At the end of all three dates, the monarch chooses which of their partners they'd like to ask on a second date. In this dataset, the ranking of a conversation's PNR (highest, lowest, or middle of that episode) significantly predicted whether that conversation resulted in a second date in a logistic GLM (coefficient=0.45, p=0.04; Fig. 4B).

Taken together, these results provide preliminary evidence that how well a potential couple jointly negotiates semantic space, as indexed by pink noise, in their initial conversation can be used to predict partner choice in a natural setting.

Pairs that engage in regular conversations over the course of several weeks demonstrate stronger pink noise signals over time

Across various non-social domains, like walking gait and time estimation, pink noise signals tend to become stronger with experience, both in the natural course of children's development and over repeated practice trials for adults (32). If pink noise in a conversation's semantic similarity trajectory represents healthy coordination, we should expect an analogous result in conversation. Specifically, we hypothesized that pairs of interlocutors who engaged in repeated conversations with each other would exhibit noise patterns that became pinker over time as they improved at the task.

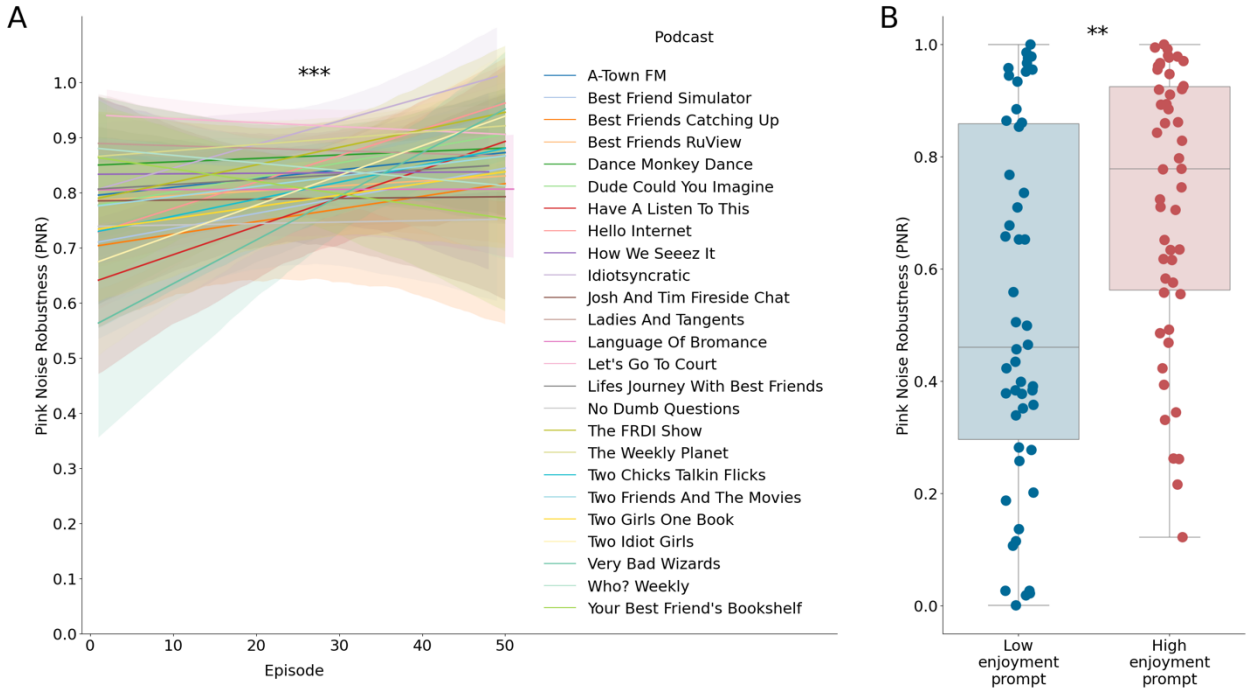

Fig. 5. Evidence of pink noise as a hallmark of optimal interpersonal semantic coordination from repeated conversations and large language model representations. (A) linear fits for the relationship between episode number and PNR for 25 dyadic conversation podcasts. Consistent with an account of pink noise representing an optimally coordinated regime, speakers trend towards pinker noise as they gain experience adapting to and with their interlocutor. Right: Distributions of PNR scores generated for conversations generated by Claude when asked to generate high- and low-enjoyment dialogues. Significance is indicated as following: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

We tested this hypothesis in the natural experiment posed by the "two dudes talking" genre of podcast. Across 25 podcasts in which two hosts have relatively unconstrained conversations (see methods for full criteria), conversation PNR increased significantly across the first 50 episodes (coefficient = 0.0018; $p<0.001$), growing by an average of 0.09, or 9% of the range of possible

PNRs (Fig. 5A). This finding suggests that pink noise reflects an optimal state of interpersonal coordination in semantic space during conversation, and that pairs trend towards this optimal state with practice.

A large language model spontaneously produces conversations with stronger pink noise signals when asked to generate good conversations

Finally, we sought to test whether a large language model (LLM), Claude (33), would spontaneously produce conversations with higher PNRs when prompted to generate "good" (vs. "bad") dyadic dialogues. If a relationship between pink noise and interaction quality exists in the large quantity of human-generated text used to train LLMs, we might expect to see that relationship mirrored in an LLM's internal representation of conversation "goodness." First, we tested whether Claude could successfully modulate the perceived enjoyability of a conversation by prompting Claude to generate enjoyable versus unpleasant dialogue scripts and asking online participants to rate these conversations on a scale of 1-5. The conversations generated by asking for enjoyable dialogues received an average score of 4.8 (SD = 0.2), while those generated by asking for unpleasant dialogues received an average score of 2.4 (SD=0.6). The distribution-level differences were significant ($t=27.1$ $p<0.001$), confirming that Claude could successfully modulate the perceived enjoyability of a dialogue. Then we tested our main hypothesis: the high-enjoyment conversations had significantly higher PNRs than their low-enjoyment counterparts (t-stat=3.3, $p<0.01$; Fig. 5B). This suggests that a system with no explicit training (to our knowledge) around the idea that cooperative interactions exhibit pink noise has nonetheless picked up on this tendency from ingesting large amounts of human-generated text.

Discussion

Good conversation, despite its ubiquity and apparent simplicity, poses a stubborn definitional challenge. Here, we demonstrate that treating conversation as a complex non-linear system allows for greater insights into how individual minds meet to form an emergent, irreducible whole and represents a viable framework for mechanistically understanding conversational success. Building on a strong theoretical tradition surrounding coordination dynamics as a core principle of social (28, 34-38), this work represents the first robust demonstration that approaches to semantics rooted in interaction-dominance can be used at scale to predict positive conversational outcomes across datasets and contexts.

In this paper, we have demonstrated that interaction-dominant mutual adaptation in high-dimensional semantic space, as indexed by pink noise, predicts positive conversational outcomes like enjoyment and connection. We first showed this in a large open conversation dataset consisting of casual conversations between strangers. We then conceptually replicated this finding in a separate dataset consisting of a qualitatively different kind of conversation: deep and potentially difficult conversations about formative life events between friends of different races. Future work is needed to develop a more nuanced understanding of how mutual adaptation plays out in a broader array of conversational contexts, with an especial focus on asymmetry and conversational goals. How might these dynamics change during political debates or tutoring sessions, and how might deviations from pink noise reflect moment-by-moment fluctuations in dominance or understanding? Future work would also benefit from longer conversations: for a

10-minute conversation, it is only possible to generate a conversation-wide measure of interaction-dominance; for an hour-long conversation, one could use a sliding window approach to identify specific techniques that interlocutors use to move a conversation into white, pink, or red noise.

We have also demonstrated that our results generalize outside of a laboratory environment. Using the natural experiment posed by speed dating podcasts, we show that mutual adaptation, as indicated by pink noise, predicts whether real-world first date conversations resulted in a second date. Additionally, we tested a longitudinal hypothesis in naturally-occurring conversations: if pink noise represents an optimal mode of coordination, interlocutors who engage in repeated conversations with each other over the course of many weeks should demonstrate stronger pink noise signals in their conversations over this period. This proved true in a collection of dyadic conversation podcasts, which is the first demonstration that a trend toward pink noise in repeated coordination behaviors holds true in a space as abstract as semantics in conversation. Beyond its novelty, this finding is important because it suggests that semantic coordination is a learnable skill.

Finally, we sought to test whether a large language model, after exposure to a large quantity of natural language including conversations, would develop entangled representations of enjoyment and mutual adaptation. Despite no explicit training on mutual adaptation as a hallmark of successful conversation, Claude produced conversations with stronger pink noise signals when asked to write enjoyable, rather than unpleasant, dialogues. This suggests that enjoyable emergence is a general property conserved in the wide array of English-language test corpora used to train large language models.

The main limitation of this work is that it is observational and correlational: we take existing free-form conversations, assay the pink noise signal, and use the strength of this signal to predict positive conversational outcome variables. This approach has an important strength: it gives us reason to believe that semantic mutual adaptation does happen in natural conversation, and therefore the predictive power it has towards positive outcomes is not limited to artificial lab-based conversation. However, if we want to develop a causal account, it will be necessary to build experimental tasks in which we manipulate the process of semantic coordination. Our longitudinal study gives us reason to believe that semantic mutual adaptation improves with experience but does not shed light on which aspects of experience are most important. Is this improvement dependent on conversing with the same interlocutor over time; does it only apply to conversations with that same interlocutor? Could we scaffold this process by explicitly teaching interlocutors about mutual adaptation, or by providing live prompts when conversations enter a period dominated by low- or high-frequency shifts in semantic space? Answering these questions will be important for using our findings to improve real-world conversations. Of particular interest is the development of behavioral supports to help scaffold more comfortable conversation for people with social deficits, like those on the autism spectrum. Additionally, building an interventional account of healthy mutual adaptation would allow for the development of more human-like artificial conversation agents, and could set the stage for technologically-mediated conversations about difficult issues, like political or racial identity.

Our approach poses a contrast – but not necessarily a contradiction – to classical theories of linguistic entrainment. Past work has repeatedly demonstrated that interlocutors tend to converge over the course of a conversation across a variety of domains: vocal pitch (39), body movements (40), lexical properties. (41), concept names (42) – even perception of one's own personality traits (43). Further, these patterns of convergence, generally termed entrainment, are largely associated with better conversation outcomes. Our findings comport with this literature – a positive slope for semantic similarity over the course of a conversation is indeed predictive of conversation enjoyment and connection. However, our framework differs from entrainment in some key ways and explains additional variance in conversation enjoyment beyond linear measures of semantic similarity. By introducing an adaptive role for moving away from each other in semantic space, interaction-dominant approaches solve entrainment's asymptote problem: what happens when speakers become perfectly synchronized? On a conceptual level, this framework also allows for individual idiosyncrasies to play an important role in the emergent behavior of a conversational system instead of being attenuated by synchronization. We present this contrast not as an argument against entrainment, but rather as a necessary corollary that embraces the full complexity of similarity dynamics in conversation.

We additionally hope that this distinction between simple convergence, or entrainment, and mutual adaptation will prove useful in the burgeoning hyperscanning literature. An ever-increasing literature in functional neuroimaging treats neural synchrony as a variable of interest, connecting this construct to individual trait differences, one's place in a social network, and more. However, neural synchrony –when defined as the correlation between two individuals' neural activity at a given timepoint-- is most useful when individuals are scanned separately while engaging in the same external stimulus. When participants interact with each other, complex behavior emerges. So too should complex neural similarity patterns.

In this paper, we have responded to a widespread call for the acknowledgement of complex interaction dynamics during conversation (34). We have demonstrated that mutual adaptation, as indexed by pink noise, explains variance in interlocutors' enjoyment and sense of connection in conversations beyond that explainable with traditional convergence-based approaches, and can predict real-world partner choice. We hope that this work will not just provide a new method for quantifying conversation and predicting outcomes, but additionally serve as a springboard for future work in communications theory and practice that acknowledges the complex dynamic nature of conversation.

Data and Code Availability

All data and materials are available via OSF (https://osf.io/7ju2s/) with the following exceptions: conversation transcripts from the CANDOR dataset must be requested from the original authors, and conversation transcripts from the Interracial Friends corpus contain personally identifying information and cannot be shared publicly. Tutorial code for generating pink noise robustness scores from conversation transcripts can be found on GitHub (https://github.com/KatieONell/pink-noise).

References

1. E. E. Levine, A. R. Roberts, T. R. Cohen, Difficult conversations: navigating the tension between honesty and benevolence. *Current Opinion in Psychology* **31**, 38–43 (2020).
2. P. Dao, Effects of task goal orientation on learner engagement in task performance. *International Review of Applied Linguistics in Language Teaching* **59**, 315–334 (2021).
3. C. N. Wright, Educational Orientation and Upward Influence: An Examination of Students' Conversations about Disappointing Grades. *Communication Education* **61**, 271–289 (2012).
4. E. M. Templeton, L. J. Chang, E. A. Reynolds, M. D. Cone LeBeaumont, T. Wheatley, Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences* **119**, e2116915119 (2022).
5. A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson, C. Fitzpatrick, T. Glazer, D. Knox, A. Liebscher, S. Marin, The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances* **9**, eadf3197 (2023).
6. D. A. Coker, J. Burgoon, The Nature of Conversational Involvement and Nonverbal Encoding Patterns. *Human Communication Research* **13**, 463–494 (1987).
7. M. G. Frank, P. Ekman, W. V. Friesen, Behavioral markers and recognizability of the smile of enjoyment. *Journal of Personality and Social Psychology* **64**, 83–93 (1993).
8. A. S. Gabay, A. Pisauro, K. C. O'Nell, M. A. Apps, Social environment-based opportunity costs dictate when people leave social interactions. *Communications Psychology* **2**, 42 (2024).
9. S. Tsang, K. Barrentine, S. Chadha, S. Oishi, A. Wood, Social exploration: How and why people seek new connections. *Psychological Review* (2024).
10. C. Geishauser, C. van Niekerk, H. Lin, N. Lubis, M. Heck, S. Feng, M. Gašić, "Dynamic Dialogue Policy for Continual Reinforcement Learning" in Proceedings of the 29th International Conference on Computational Linguistics, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na, Eds. (International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022; https://aclanthology.org/2022.coling-1.21), pp. 266–284.
11. P. A. Heeman, "Representing the Reinforcement Learning state in a negotiation dialogue" in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding* (2009; https://ieeexplore.ieee.org/document/5373413), pp. 450–455.
12. M. Gallotti, C. D. Frith, Social cognition in the we-mode. *Trends in Cognitive Sciences* **17**, 160–165 (2013).
13. S. Wallot, D. G. Kelty-Stephen, Interaction-Dominant Causation in Mind and Brain, and Its Implication for Questions of Generalization and Replication. *Minds & Machines* **28**, 353–374 (2018).
14. M. Usher, M. Stemmler, Z. Olami, Dynamic Pattern Formation Leads to 1/f Noise in Neural Populations. *Phys. Rev. Lett.* **74**, 326–329 (1995).

15. D. G. Stephen, D. Mirman, Interactions Dominate the Dynamics of Visual Cognition. *Cognition* **115**, 154 (2010).
16. P. Bak, C. Tang, K. Wiesenfeld, Self-organized criticality: An explanation of the 1/ *f* noise. *Phys. Rev. Lett.* **59**, 381–384 (1987).
17. J. J. Collins, C. J. De Luca, Upright, correlated random walks: A statistical-biomechanics approach to the human postural control system. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **5**, 57–63 (1995).
18. C.-K. Peng, S. Havlin, J. M. Hausdorff, J. E. Mietus, H. E. Stanley, A. L. Goldberger, Fractal mechanisms and heart rate dynamics: Long-range correlations and their breakdown with disease. *Journal of Electrocardiology* **28**, 59–65 (1995).
19. K. Linkenkaer-Hansen, S. Monto, H. Rytsälä, K. Suominen, E. Isometsä, S. Kähkönen, Breakdown of Long-Range Temporal Correlations in Theta Oscillations in Patients with Major Depressive Disorder. *The Journal of Neuroscience* **25**, 10131 (2005).
20. D. H. Abney, A. Paxton, R. Dale, C. T. Kello, Movement dynamics reflect a functional role for weak coupling and role structure in dyadic problem solving. *Cognitive Processing* **16**, 325–332 (2015).
21. O. Mayo, I. Gordon, In and out of synchrony—Behavioral and physiological dynamics of dyadic interpersonal coordination. *Psychophysiology* **57**, e13574 (2020).
22. R. Dale, M. J. Spivey, Unraveling the Dyad: Using Recurrence Analysis to Explore Patterns of Syntactic Coordination Between Children and Caregivers in Conversation. *Language Learning* **56**, 391–430 (2006).
23. H. P. Branigan, M. J. Pickering, A. A. Cleland, Syntactic co-ordination in dialogue. *Cognition* **75**, B13–B25 (2000).
24. R. Fusaroli, K. Tylén, Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance. *Cognitive Science* **40**, 145–171 (2016).
25. S. Schneider, A. G. Ramirez-Aristizabal, C. Gavilan, C. T. Kello, Complexity matching and lexical matching in monolingual and bilingual conversations. *Bilingualism: Language and Cognition* **23**, 845–857 (2020).
26. D. H. Abney, A. Paxton, R. Dale, C. T. Kello, Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General* **143**, 2304–2315 (2014).
27. P. J. Taylor, S. Thomas, Linguistic Style Matching and Negotiation Outcome. *Negotiation and Conflict Management Research* **1** (2008).
28. D. Angus, Recurrence Methods for Communication Data, Reflecting on 20 Years of Progress. *Front. Appl. Math. Stat.* **5** (2019).
29. M. T. Tolston, M. A. Riley, V. Mancuso, V. Finomore, G. J. Funke, Beyond frequency counts: Novel conceptual recurrence analysis metrics to index semantic coordination in team communications. *Behav Res Methods* **51**, 342–360 (2019).
30. C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49**, 1685–1689 (1994).

31. L. Rydin Gorjão, G. Hassan, J. Kurths, D. Witthaut, MFDFA: Efficient multifractal detrended fluctuation analysis in python. *Computer Physics Communications* **273**, 108254 (2022).
32. G. C. Van Orden, H. Kloos, S. Wallot, "Living in the Pink: Intentionality, Wellbeing, and Complexity" in *Philosophy of Complex Systems*, C. Hooker, Ed. (North-Holland, Amsterdam, 2011; https://www.sciencedirect.com/science/article/pii/B9780444520760500225)vol. 10 of *Handbook of the Philosophy of Science*, pp. 629–672.
33. Anthropic, Claude 3.7 Sonnet. https://www.anthropic.com/claude/sonnet.
34. U. Hasson, C. D. Frith, Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150366 (2016).
35. T. Wheatley, M. A. Thornton, A. Stolk, L. J. Chang, The Emerging Science of Interacting Minds. *Perspect Psychol Sci* **19**, 355–373 (2024).
36. R. Fusaroli, J. Rączaszek-Leonardi, K. Tylén, Dialog as interpersonal synergy. *New Ideas in Psychology* **32**, 147–157 (2014).
37. E. Tognoli, M. Zhang, A. Fuchs, C. Beetle, J. A. S. Kelso, Coordination Dynamics: A Foundation for Understanding Social Behavior. *Front. Hum. Neurosci.* **14** (2020).
38. I. Ravreby, Y. Shilat, Y. Yeshurun, Liking as a balance between synchronization, complexity and novelty. *Sci Rep* **12**, 3181 (2022).
39. J. Michalsky, H. Schoormann, "Pitch Convergence as an Effect of Perceived Attractiveness and Likability" (2017; https://www.isca-archive.org/interspeech_2017/michalsky17_interspeech.html), pp. 2253–2256.
40. J. L. Lakin, T. L. Chartrand, Using Nonconscious Behavioral Mimicry to Create Affiliation and Rapport. *Psychol Sci* **14**, 334–339 (2003).
41. S. E. Brennan, H. H. Clark, Conceptual pacts and lexical choice in conversation. *J Exp Psychol Learn Mem Cogn* **22**, 1482–1493 (1996).
42. S. E. Brennan, Lexical Entrainment in Spontaneous Dialog. (1996).
43. C. Welker, T. Wheatley, G. Cason, C. Gorman, M. Meyer, Self-views converge during enjoyable conversations. *Proc Natl Acad Sci U S A* **121**, e2321652121 (2024).